A Style-Transfer-Based Augmentation Framework for Ship Recognition in Infrared Imagery

Yan-Ann Chen[1], Hung-Wen Lin[2]*

[1] *Department of Electronic Engineering, National Taipei University of Technology, Taiwan*
[2] *Department of Computer Science and Engineering, Yuan Ze University, Taiwan*
chenya@ntut.edu.tw, s1106108@mail.yzu.edu.tw

Abstract

Maritime security relies on reliable ship recognition, where infrared (IR) imagery is crucial for nighttime and adverse weather monitoring. However, the low resolution and scarcity of labeled IR data severely limit recognition accuracy. This study proposes a multi-stage augmentation framework that integrates style transfer, conditional generative modeling, and class-specific augmentation guided by large language models. Experiments under the IR-only setting show significant gains: for VGG16, overall accuracy improves by 1.4 times, from 56.61% to 79.80%, and nighttime IR accuracy improves by 1.6 times, from 44.81% to 71.43%. These results confirm the effectiveness of the proposed framework and its potential for vision tasks constrained by limited and low-quality data.

Keywords: Data scarcity mitigation, Generative adversarial networks, Infrared ship recognition, Maritime surveillance, Style-transfer augmentation

1 Introduction

With the rapid growth of international maritime activities, ensuring safety and operational stability has become a global priority. Maritime regions of strategic importance are increasingly exposed to challenges such as smuggling, illegal fishing, human trafficking, and even military operations that elevate defense and surveillance demands. In this context, accurate ship recognition is essential for supporting both civilian and defense applications, enabling timely responses and improving situational awareness. In recent years, the advancement of artificial intelligence has driven the development of intelligent surveillance systems, yet infrared (IR) imagery, which captures heat distribution through infrared sensors, remains difficult to utilize effectively due to data scarcity and low resolution, particularly at night.

Maritime surveillance traditionally relies on visible-light cameras and radar systems, yet both face critical limitations under low-light or adverse weather conditions. Visible-light cameras lose effectiveness at night, while radar, although robust in range and weather, produces low-resolution imagery insufficient for fine-grained ship recognition. Infrared sensors offer a promising alternative because they operate reliably in darkness and poor weather.

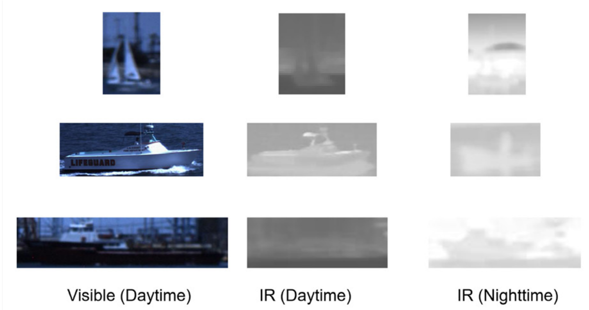


Figure 1. Examples of visible-light and infrared ship images under daytime and nighttime conditions (Samples are taken from the VAIS dataset [1].)

Figure 1 shows examples of visible-light and infrared ship images under daytime and nighttime conditions. However, IR images are typically of low resolution, and nighttime IR data are particularly scarce. For example, the VAIS dataset provides paired visible–daytime IR samples, but nighttime IR images are sparse and unpaired. This imbalance hinders the development of AI-based recognition models, which demand large and diverse datasets to achieve robust performance.

To overcome these challenges, this study introduces a style-transfer-based data augmentation framework designed to strengthen recognition in IR imagery. The framework leverages visible-light images, which are more abundant and of higher quality, and transforms them into IR-style counterparts using style-transfer methods. Furthermore, generative adversarial networks (GANs) are employed to generate additional synthetic IR images, increasing both the volume and diversity of training data. In addition, recent advances in generative models, including the use of large language models (LLMs) for guiding or supplementing image synthesis, further expand dataset diversity and compensate for class imbalance. By enriching the dataset in this way, the proposed approach alleviates the shortage of nighttime IR imagery and enhances the accuracy of deep learning recognition models, with particular benefits for nighttime ship recognition.

The major contributions of this study are summarized as follows:

*Corresponding Author: Yan-Ann Chen; Email: chenya@ntut.edu.tw
DOI: https://doi.org/10.70003/160792642026092705015

(1) We propose a multi-stage augmentation pipeline that leverages CycleGAN [2] and Pix2Pix [3] to translate visible images into IR-style counterparts, effectively increasing the availability of IR-like training data.

(2) A Conditional Deep Convolutional GAN (CDCGAN) [4-5] is employed to synthesize class-controllable IR images, addressing class imbalance and enhancing dataset diversity.

(3) Integration with recognition models: By incorporating both style-transferred and generative IR samples into recognition model training (CNN and VGG16), the proposed framework achieves substantial performance improvements, especially in nighttime IR recognition tasks.

(4) Practical relevance: The framework demonstrates the feasibility of using AI-based augmentation to mitigate the shortage of IR imagery, thereby contributing to more robust and reliable maritime monitoring systems.

2 Related Work

Infrared Imaging. Infrared imaging is fundamentally based on Planck's law of black-body radiation, which describes the spectrum of electromagnetic radiation emitted by an object at a given temperature. Since all objects above absolute zero radiate energy, the captured IR intensity distribution directly corresponds to surface temperature patterns. According to wavelength, IR radiation is categorized into short-wave infrared (SWIR), mid-wave infrared (MWIR), and long-wave infrared (LWIR). Unlike visible-light images, IR images are typically rendered in grayscale because they reflect thermal characteristics rather than color information.

In the maritime domain, the VAIS (Visible and Infrared Ship) dataset [1] provides the first publicly available paired collection of visible-light (RGB) and long-wave infrared (LWIR) ship images. The dataset was acquired using a Sofradir-EC Atom 1024 camera, which was among the highest-resolution LWIR sensors available at that time. It includes six ship categories: cargo, medium-other, passenger, sailing, small, and tugboats. Recognition experiments using VGG16 achieved 54% accuracy for daytime LWIR images and 59.9% for nighttime LWIR images, highlighting both the potential and the challenges of IR-based ship recognition.

In addition, multispectral datasets have also been developed in other domains. The KAIST multispectral pedestrian dataset [6] provides a large-scale collection of paired color and IR images for pedestrian detection under varying illumination. Although its content is not maritime-related, the paired structure of the dataset allows it to be leveraged for training style-transfer models, thereby supporting our augmentation framework.

Applications of Infrared Imaging. IR imaging has emerged as a critical sensing modality in scenarios where visible-light imaging is unreliable. Unlike conventional cameras that depend on reflected light, IR sensors capture thermal radiation, allowing continuous operation at night and under adverse weather. This capability makes IR particularly valuable for both large and small object detection in complex environments, such as land battlefields and surveillance systems, where targets may be camouflaged or have weak signatures. Recent studies have shown that deep learning methods, including hierarchical context fusion networks for small infrared objects [7] and enhanced YOLO-based detectors [8], can effectively mitigate challenges such as background clutter and low-resolution inputs. These advances demonstrate that IR imaging not only improves robustness in defense and military monitoring [9] but also has broad potential for civilian applications requiring reliable perception in low-light conditions.

For maritime scenarios, IR imaging has been increasingly adopted for ship detection and recognition because of its ability to penetrate atmospheric disturbances such as fog and haze, enabling passive and persistent observation during both daytime and nighttime. Recent studies have focused on enhancing detection and classification performance in IR ship imagery. For example, STFA-YOLO [10] improved ship detection in complex marine environments, achieving real-time performance but emphasizing detection rather than fine-grained recognition. Reference [11] analyzed recognition performance across multi-band IR images, showing the impact of spectral variations on classification accuracy, although the scarcity of nighttime IR data remained a challenge. Reference [12] introduced an infrared target recognition approach based on fuzzy comprehensive evaluation to mitigate noise and blurred edges, yet its reliance on handcrafted features limited adaptability to diverse conditions. In contrast to these approaches, our work emphasizes a style-transfer-based augmentation framework that exploits abundant visible-light images together with generative models to synthesize IR samples. This strategy is particularly aimed at compensating for the limited availability of nighttime IR data, thereby advancing fine-grained ship recognition in challenging maritime environments.

Style-Transfer-Based Data Augmentation. Traditional augmentation methods, such as rotation, flipping, or noise injection, are widely applied to increase image diversity [13]. However, these operations are less meaningful for ship imagery that is already cropped and aligned, and may even distort structural features. This has motivated the adoption of style-transfer techniques based on GANs to synthesize IR-style data from visible-light images.

In this direction, reference [14] introduced ThermalGAN, an improved Pix2Pix variant trained on the VAIS dataset. While their results showed an approximate 8% gain in VGG-F accuracy when ThermalGAN-generated samples were added, the method relied heavily on parameter tuning and exhibited limited improvements for nighttime IR data, where scarcity remained critical. More recently, reference [15] proposed an enhanced StyleGAN2 framework with a variational auto-encoder mapping network and self-attention to generate higher-quality IR ship images. Although this approach improved image fidelity and diversity, its primary focus was on visual quality and lightweight detection, leaving the quantitative link between augmentation and recognition performance less explored.

Compared with these works, our work explicitly positions style-transfer and conditional GAN–based augmentation as the central mechanism to expand training diversity. Instead of primarily optimizing architectures or visual fidelity, our framework systematically evaluates how different augmentation strategies impact recognition accuracy, with particular emphasis on improving performance under challenging nighttime IR conditions.

3 Framework Design

3.1 Problem Formulation

In maritime surveillance, visible imagery (x^v) is abundant and of relatively high resolution, whereas infrared imagery (x^t) is both scarce and of lower resolution. Samples may appear as paired visible-infrared observations or as unpaired instances containing only one modality. In particular, nighttime IR imagery (X_t^{night}) is especially limited, which severely constrains the training of recognition models. We denote the labeled dataset as $D = \{(x_i^v, x_i^t, y_i)\}_{i=1}^{N}$, where $x_i^v \in X_v$ represents a visible image, $x_i^t \in X_t$ represents an IR image that may be absent, and $y_i \in Y$ is the ship category label.

The objective is to improve the recognition performance across the entire dataset D, with particular emphasis on the scarce and challenging subset X_t^{night}. Therefore, we aim to learn a recognition model f_θ: $X \rightarrow Y$ with $X = X_v \cup X_t$, where θ are the model parameters, under the constraint that IR data, especially at night, are limited in both quantity and quality.

To address this limitation, we introduce an augmentation pipeline G, trained on available visible and IR samples, that generates synthetic IR-style images either conditioned on visible inputs or on class-level information. We denote generated samples as $\tilde{x}_j^t = G(s_j)$, where $s_j \in \{x_j^v, (z, c)\}$ and (z, c) denotes latent noise with conditional labels. The augmented training set is then $D' = D \cup \{(\tilde{x}_j^t, y_j)\}_{j=1}^{M}$, where M is the number of generated samples.

The final goal is to train f_θ on D' such that recognition accuracy is maximized, which can be expressed as: $\theta^* = argmax_\theta\, Pr_{(x,y)\sim D_{test}}[f_\theta(x) = y]$, where D_{test} denotes the test distribution drawn from real data. Here, Pr[·] represents the probability of correct classification, i.e., the expected accuracy of f_θ on unseen samples. Performance improvements on IR imagery, particularly X_t^{night}, directly contribute to enhanced recognition over the entire dataset.

3.2 Augmentation Framework

The proposed framework addresses the scarcity and low resolution of IR imagery by progressively enriching the training dataset through style-transfer and generative augmentation. As illustrated in Figure 2, the framework is organized into four sequential stages: (1) style-transfer augmentation, (2) conditional generative augmentation, (3) class-specific augmentation, and (4) recognition model training. Together, these stages form a unified pipeline that systematically expands both the quantity and diversity of IR training data, ultimately improving recognition performance under challenging scenarios.

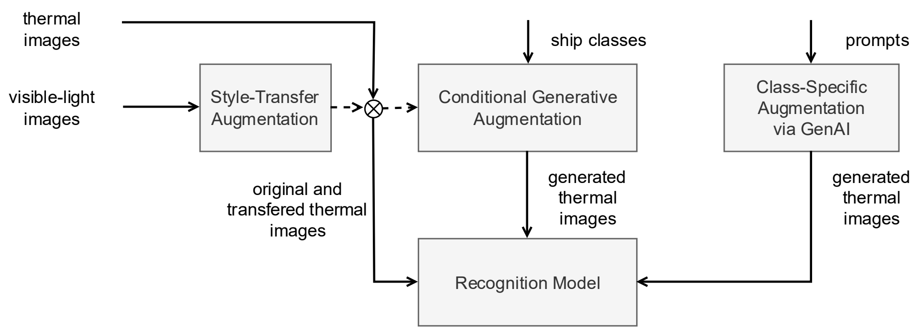


Figure 2. Overview of the proposed augmentation pipeline combining style-transfer, conditional generative modeling, and class-specific generative AI for improving ship recognition in IR imagery

Stage 1: Style-Transfer Augmentation. Visible-light images are first translated into IR-style images using both unpaired and paired image-to-image translation methods. In particular, CycleGAN [2] is employed for unpaired translation, while Pix2Pix [3] is used when paired samples are available. This stage leverages the abundance of visible imagery to generate synthetic IR counterparts, thereby alleviating the limited availability of IR data.

Both CycleGAN and Pix2Pix rely on adversarial training between a generator G and a discriminator D. The general adversarial loss is formulated as

$$\begin{aligned}&\mathcal{L}_{GAN}(G,D)\\&=E_{y\sim p_{data}(y)}\left[\log D(y)\right]\\&+E_{x\sim p_{data}(x)}\left[\log\left(1-D\left(G(x)\right)\right)\right],\end{aligned}$$

where $p_{data}(x)$ and $p_{data}(y)$ denote the distributions of visible and IR images, respectively.

The distinction lies in the additional constraints used. CycleGAN introduces a cycle-consistency loss to preserve semantic content during domain translation:

$$\mathcal{L}_{cyc}\left(G_X, G_Y\right) = E_{x\sim p_{data}(x)}\left[\left\|G_Y\left(G_X\left(x\right)\right)-x\right\|_1\right]+ E_{y\sim p_{data}(y)}\left[\left\|G_X\left(G_Y\left(y\right)\right)-y\right\|_1\right]$$

Pix2Pix, on the other hand, leverages paired samples and incorporates an L_1 reconstruction term:

$$\mathcal{L}_{L1}\left(G\right) = E_{x,y}\left[\left\|y-G\left(x\right)\right\|_1\right]$$

to ensure both realism and fidelity in the translated images.

In practice, style-transfer models can be trained on different visible-infrared domains, producing complementary translation variants that further enrich the diversity of synthetic IR images for subsequent augmentation stages.

Stage 2: Conditional Generative Augmentation. In the second stage, both real IR images and style-transferred images from Stage 1 are leveraged to train a conditional generative model, as illustrated in Figure 2. The dashed line indicates the inputs used during training, while the solid arrows represent the conditional information provided at generation time. By conditioning on ship classes, the generator is able to produce IR images that are not only realistic but also aligned with specific categories, thereby enriching the dataset and mitigating class imbalance. To this end, a Conditional Deep Convolutional GAN (CDCGAN) is employed, since its convolutional architecture effectively captures spatial features and enhances the quality of generated IR images.

Formally, given a noise vector z and a conditional label c, the generator learns a mapping G: $(z, c) \mapsto \tilde{x}$, where the symbol $\mapsto$ denotes that the input pair (z, c) is transformed by the generator G into a synthetic IR image $\tilde{x}$. The discriminator, in turn, evaluates image–condition pairs as D: $(x, c) \mapsto [0, 1]$, where the output corresponds to the probability that x is a real IR image under condition c. The learning process is guided by a conditional adversarial loss:

$$\mathcal{L}_{cGAN}\left(G, D\right) = E_{x,c}\left[\log D\left(x,c\right)\right] + E_{z,c}\left[\log\left(1-D\left(G\left(z,c\right),c\right)\right)\right]$$

Similar to Stage 1, the framework is not restricted to a single dataset or label configuration. In practice, multiple class-conditional generative models can be trained on different domains, producing complementary synthetic variants that further enhance the diversity and balance of IR imagery for subsequent recognition.

Stage 3: Class-Specific Augmentation with Large Generative Models. To further address category imbalance, recognition performance is monitored at the class level via validation data. Classes exhibiting low recognition accuracy are selectively augmented with the help of large generative models (e.g., LLM-guided image generators). Unlike CDCGAN in Stage 2, which primarily balances the overall class distribution, this stage targets categories that remain underperforming, especially in nighttime IR conditions, by generating more diverse and semantically rich samples.

For example, an LLM-based image generator can be prompted with task-specific queries such as:

"Generate IR-style nighttime images of a tugboat captured by an infrared camera."

This guided augmentation increases intra-class diversity for hard-to-recognize categories, thereby improving recognition robustness in the final model training.

Stage 4: Recognition Model Training. In the final stage, all IR images, including both the original data and the augmented samples generated in Stages 1–3, are consolidated to form a comprehensive training set. The recognition model is trained on this enriched dataset, allowing it to leverage the expanded volume and diversity of IR imagery. By integrating style-transfer augmentation, conditional generative augmentation, and class-specific generative augmentation, the framework ensures that the recognition model is exposed to a wide range of intra-class variations, thereby improving robustness and accuracy in IR ship recognition.

Given an input image x with label $y \in \{1, ..., C\}$, the recognition model outputs a probability distribution $\hat{y} = f_\theta(\mathrm{x})$ over C classes. The training objective minimizes the standard cross-entropy loss:

$$\mathcal{L}_{cls}\left(\theta\right) = -E_{(x,y)\sim\mathbb{D}'}\sum_{c=1}^{C} y_c \log \widehat{y_c}$$

where the expectation is taken over the augmented dataset D', which combines the original dataset with generated IR-style images. Here, y_c denotes the one-hot encoding of the true label and $\hat{y}_c$ represents the predicted probability of class c.

4 Experimental Results

4.1 Dataset

Two datasets are employed in this study with distinct roles. The VAIS dataset [1] contains 2,865 ship images, including 1,623 visible-light (RGB) and 1,242 infrared (IR) images, with 1,088 paired samples. Following the original split, 1,412 images are used for training and 1,453 for testing, among which 154 are nighttime IR images. VAIS serves as the benchmark dataset for ship recognition experiments.

The KAIST dataset [6] provides 95,324 paired pedestrian images (color–thermal), with 50,184 pairs for training and 45,140 for testing. Although it is not maritime-oriented, KAIST is exploited to pre-train style-transfer models (CycleGAN and Pix2Pix), leveraging its scale to improve visible-to-IR translation in the proposed framework.

Since the VAIS images are of variable resolution and the KAIST images have a fixed resolution of 640×512, all samples were resized to 224×224 for consistency with CNN and VGG16 input requirements. Visible-light images are treated as three-channel RGB, while IR images are replicated across three channels to form pseudo-RGB representations, ensuring compatibility with the recognition models.

4.2 Experimental Setup

All experiments were conducted on an NVIDIA Tesla T4 GPU (Google Colab platform). This platform was chosen for its accessibility and reproducibility, while the T4 GPU provided sufficient computational power to train generative and recognition models within practical time limits.

The experimental framework involved three categories of models: (1) style-transfer models, including CycleGAN and Pix2Pix; (2) augmentation models, specifically a Conditional Deep Convolutional GAN (CDCGAN); and (3) recognition models, including a lightweight CNN and VGG16.

CycleGAN. CycleGAN was implemented with ResNet-based generators containing residual blocks, and PatchGAN [3] discriminators. Instance Normalization was used instead of Batch Normalization to improve style-transfer fidelity, especially for local textures. Training followed the Adam optimizer and included cycle-consistency and identity constraints to stabilize unpaired translation.

Pix2Pix. The Pix2Pix generator adopted a U-Net [16] encoder–decoder with skip connections to preserve spatial details. Its discriminator was a PatchGAN, which enforces local-level realism by operating on small image patches. Batch Normalization was applied throughout the network, and training was carried out with Adam, combining adversarial training with an L1 reconstruction objective for paired translation.

CDCGAN. The CDCGAN generator used transposed convolutions to upsample latent noise vectors concatenated with class labels, producing class-conditional synthetic IR images. The discriminator, implemented as a convolutional network, evaluated image–label pairs to distinguish real from synthetic samples. Batch Normalization layers were applied across both generator and discriminator to stabilize training, and adversarial optimization was performed with binary cross-entropy loss.

Recognition Models. Two models were employed to evaluate the impact of augmentation. The lightweight CNN consisted of five layers of convolution, activation, and max-pooling, followed by two fully connected layers with dropout and a softmax classifier. This compact design balances accuracy and efficiency, making it suitable for resource-constrained deployment. The VGG16 model was initialized with ImageNet-pretrained weights. Its final classification layer was replaced to match the number of ship categories, and fine-tuning was applied to the last convolutional block and fully connected layers, while earlier layers remained frozen to preserve general feature representations.

4.3 Ablation Study

We report the recognition accuracy obtained using the augmentation framework (stages 1 and 2). Two training configurations were considered:

- All (Mixed Training): Models were trained on a mixture of visible and IR images (VAIS dataset), and tested on both visible and IR subsets (daytime and nighttime). This setting evaluates the overall performance of the ship recognition.
- All IR (IR-only Training): Models were trained solely on IR images and tested on IR subsets. This setting directly evaluates the effect of style-transfer augmentation on scarce IR data.

Both CNN and VGG16 architectures were evaluated. In each case, the baseline corresponds to training without augmentation, while the augmentation settings include images generated by CycleGAN/Pix2Pix (stage 1) and CDCGAN (stage 2).

As summarized in Table 1, style-transfer augmentation substantially improved recognition accuracy across both models and training settings. For instance, CNN performance on nighttime IR images improved from 16.23% to 70.13% under the IR-only setting. Similarly, VGG16 achieved 79.8% overall accuracy when trained with augmented IR data, compared to 56.61% without augmentation. Note that results incorporating LLM-generated images are not included in this table, since they target specific underrepresented classes and may not consistently improve overall accuracy.

Table 1. Overall recognition performance by style-transfer augmentations

Method	All	Daytime	Nighttime
	CNN		
Raw VAIS	53.34%	59.56%	31.17%
Style-Transfer CDCGAN	79.94%	82.51%	70.78%
	VGG16		
Raw VAIS	59.89%	63.02%	48.70%
Style-Transfer CDCGAN	77.24%	79.23%	70.13%

Method	All IR	Daytime IR	Nighttime IR
	CNN		
Raw VAIS	51.49%	61.38%	16.23%
Style-Transfer CDCGAN	78.81%	81.60%	70.13%
	VGG16		
Raw VAIS	56.61%	60.47%	44.81%
Style-Transfer CDCGAN	79.80%	82.15%	71.43%

Effect of Style-Transfer CycleGAN/Pix2Pix. As the baseline, the Raw VAIS dataset refers to the original VAIS infrared images without any additional augmentation. To evaluate the contribution of stage 1, three augmentation settings are considered: Style-Transfer (CycleGAN), Style-Transfer (Pix2Pix), and Style-Transfer (Both).

Representative examples are illustrated in Figure 3, Figure 4, Figure 5, and Figure 6. These results show that CycleGAN captures domain-level style translation, producing IR-style counterparts from visible inputs, whereas Pix2Pix generates sharper and more structurally aligned conversions by leveraging paired data. Together, these diverse translation outputs provide complementary synthetic samples that serve as a foundation for stage 2 generative augmentation.

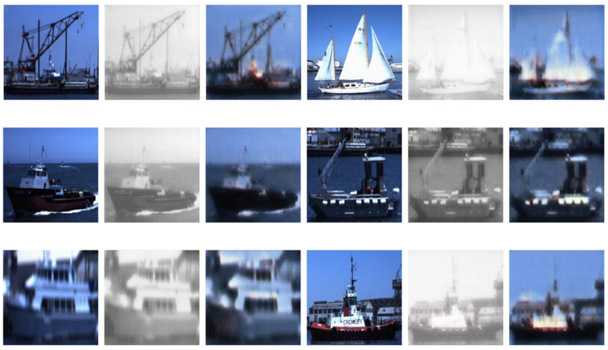


Figure 3. Examples of CycleGAN translations on the VAIS dataset [Each row contains two groups of samples, where each group consists of the original visible image (left), the translated infrared-style image (middle), and the reconstructed visible image (right).]

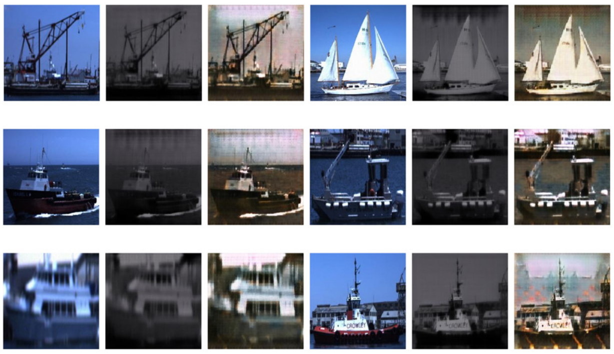


Figure 4. Examples of CycleGAN translations on the KAIST dataset [Each row contains two groups of samples, where each group consists of the original visible image (left), the translated infrared-style image (middle), and the reconstructed visible image (right).]

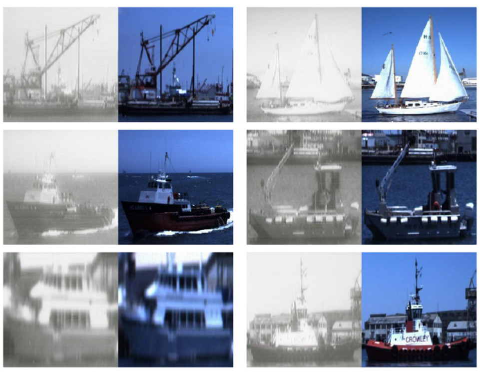


Figure 5. Examples of Pix2Pix translations on the VAIS dataset [Each row contains two groups of samples, where each group shows the translated infrared-style image (left) and its corresponding original visible image (right).]

Figure 6. Examples of Pix2Pix translations on the KAIST dataset [Each row contains two groups of samples, where each group shows the translated infrared-style image (left) and its corresponding original visible image (right).]

Quantitative results are summarized in Table 2. For CNN, recognition accuracy improved from 51.49% (baseline) to 61.17% overall and from 16.23% to 51.95% on nighttime IR images. For VGG16, accuracy increased from 56.61% to 66.57% overall and from 44.81% to 59.09% on nighttime IR images. Although the gains are modest, they demonstrate that style-transfer augmentation can alleviate data scarcity and improve recognition robustness, particularly under nighttime conditions where IR data are most limited.

The number of generated images depends on the design of the style-transfer settings. For Style-Transfer (CycleGAN), a total of 1,746 translated IR images were obtained, with 873 generated from VAIS-trained models and another 873 from KAIST-trained models. Similarly, Style-Transfer (Pix2Pix) produced 1,746 IR images, as both VAIS and KAIST paired data were exploited. Finally, Style-Transfer (Both) combined the outputs of the two methods and further applied Pix2Pix to the 1,746 visible images translated back by CycleGAN, resulting in an augmented set of 1,746×3 images. This design highlights that while both CycleGAN and Pix2Pix can independently scale to similar sizes, their combination yields the largest augmentation set by reusing translated domains.

Table 2. Recognition accuracy (IR-only) across augmentation stages

Method	All IR	Daytime IR	Nighttime IR
	CNN		
Raw VAIS	51.49%	61.38%	16.23%
Style-Transfer (CycleGAN)	50.21%	57.74%	23.38%
Style-Transfer (Pix2Pix)	53.06%	60.11%	27.92%
Style-Transfer (Both)	61.17%	63.75%	51.95%
Style-Transfer CDCGAN	**78.81%**	**81.60%**	**70.13%**
	VGG16		
Raw VAIS	56.61%	60.47%	44.81%
Style-Transfer (CycleGAN)	60.17%	65.03%	42.86%
Style-Transfer (Pix2Pix)	54.20%	56.47%	46.10%
Style-Transfer (Both)	66.57%	68.67%	59.09%
Style-Transfer CDCGAN	**79.80%**	**82.15%**	**71.43%**

Effect of Style-Transfer CDCGAN. It is important to note that the results here, as well as in "Effect of Style-Transfer CycleGAN/Pix2Pix.", are based on the IR-only setting. We focus on IR images because they represent the most challenging and data-scarce modality, and improvements here directly demonstrate the benefit of style-transfer augmentation.

Therefore, stage 2 integrates real IR images and stage 1 style-transfer outputs to train a conditional generative model. In this work, we adopt Style-Transfer (CDCGAN), where generated samples are conditioned on class labels to provide more balanced and diverse IR data. Specifically, 300 images were generated for each of the six ship classes, resulting in a total of 1,800 synthetic IR images. Representative results are shown in Figure 7.

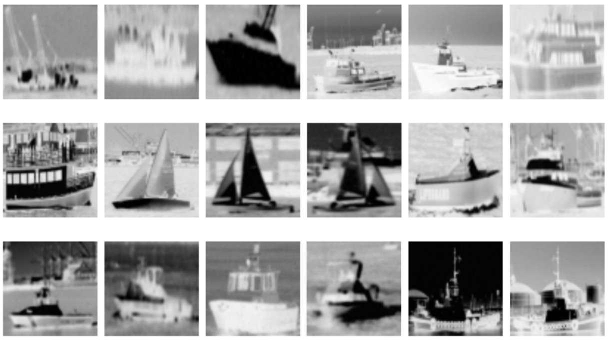


Figure 7. Infrared-style ship images generated by CDCGAN trained on augmented data from style-transfer methods

As summarized in Table 2, incorporating CDCGAN-generated samples (which already include stage 1 style-transfer outputs) leads to clear performance gains. For CNN, overall accuracy improves from 61.17% (Style-Transfer (Both)) to 78.81% (Style-Transfer CDCGAN), with nighttime IR rising from 51.95% to 70.13%. For VGG16, accuracy increases from 66.57% to 79.8% overall, and from 59.09% to 71.43% on nighttime IR. These results demonstrate that conditional generative augmentation not only strengthens recognition accuracy but also alleviates class imbalance, particularly under challenging nighttime IR conditions.

Effect of Class-Specific Augmentation with LLMs. In stage 3, we illustrate class-specific augmentation using the tugboat category as a representative minority class. We employed LLM-based prompting to generate additional samples for the tugboat class, which had the fewest training instances, as shown in Figure 8. Table 3 reports the recognition performance under the mixed training configuration (All). This setting was adopted to ensure that class-specific augmentation is evaluated in a comprehensive training context, rather than being limited to IR-only analysis as in the previous sections. The objective here is to demonstrate the feasibility and potential benefit of LLM-guided data generation when applied to scarce categories.

Figure 8. Tugboat images generated by Google Gemini to enhance class-specific diversity in the dataset

As shown in Table 3, the inclusion of LLM-generated tugboat images improves recognition accuracy on this class, leading to gains in overall performance as well. For CNN, accuracy increases from 53.6% to 59.4% (+GenAI) overall and from 46.9% to 57.1% at nighttime. For VGG16, overall accuracy improves from 56.5% to 60.9%, with nighttime performance rising from 49.0% to 55.1%. These results highlight that targeted augmentation can be particularly effective in alleviating class imbalance, demonstrating the potential of LLM-guided data generation for improving recognition in rare categories.

Table 3. Effect of generated AI-based class-specific augmentation (Tugboat class)

Method	All	Daytime	Nighttime
	CNN		
Style-Transfer CDCGAN	53.6%	70%	46.9%
+GenAI	59.4%	65%	57.1%
	VGG16		
Style-Transfer CDCGAN	56.5%	75%	49%
+GenAI	60.9%	75%	55.1%

5 Conclusion

This study proposed a style-transfer-based augmentation framework for scarce, low-resolution infrared ship imagery. By combining image translation, conditional image generation, and targeted class augmentation, the framework improved VGG16 accuracy from 56.61% to 79.80% under the IR-only setting and from 44.81% to 71.43% for nighttime IR images. The tugboat case further confirms the value of targeted augmentation for minority classes.

Beyond these improvements, the framework shows broader potential for data-scarce domains such as medical imaging, surveillance, and remote sensing. By combining style transfer with conditional generative modeling, it offers a scalable way to improve recognition under challenging nighttime IR conditions and other low-resource environments.

Acknowledgment

Y.-A. Chen's research is co-sponsored by MOST 107-2218-E-155-007-MY3, Taiwan, and MOST 110-2221-E-155-022-MY3, Taiwan.

References

[1] M. M. Zhang, J. Choi, K. Daniilidis, M. T. Wolf, C. Kanan, VAIS: A dataset for recognizing maritime imagery in the visible and infrared spectrums, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Boston, MA, USA, 2015, pp. 10-16. https://doi.org/10.1109/CVPRW.2015.7301291

[2] J. Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2223-2232.

[3] P. Isola, J. Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 1125-1134.

[4] M. Mirza, S. Osindero, Conditional generative adversarial nets, *arXiv preprint*, arXiv:1411.1784, November, 2014. https://arxiv.org/abs/1411.1784

[5] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, *arXiv preprint*, arXiv:1511.06434, November, 2015. https://arxiv.org/abs/1511.06434

[6] S. Hwang, J. Park, N. Kim, Y. Choi, I. S. Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 1037-1045. https://doi.org/10.1109/CVPR.2015.7298706

[7] S. Xu, S. Zheng, W. Xu, R. Xu, C. Wang, J. Zhang, X. Teng, A. Li, G. Li, Hcf-net: Hierarchical context fusion network for infrared small object detection, *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Niagara Falls, ON, Canada, 2024, pp. 1-6. https://doi.org/10.1109/ICME57554.2024.10687431

[8] S. Li, Y. Li, Y. Li, M. Li, X. Xu, YOLO-FIRI: Improved YOLOv5 for infrared image object detection, *IEEE Access*, Vol. 9, pp. 141861-141875, October, 2021. https://doi.org/10.1109/ACCESS.2021.3120870

[9] G. Zheng, X. Wu, Y. Hu, X. Liu, Object detection for low-resolution infrared image in land battlefield based on deep learning, *Proceedings of Chinese Control Conference (CCC)*, Guangzhou, China, 2019, pp. 8649-8652. https://doi.org/10.23919/ChiCC.2019.8866344

[10] T. Li, Y. Hu, N. Wen, STFA-YOLO: An Algorithm for Enhanced Infrared Ship Detection in Complex Marine Environments, *Proceedings of International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, Compiègne, France, 2025, pp. 2196-2201. https://doi.org/10.1109/CSCWD64889.2025.11033360

[11] Y. Li, D. Zhang, L. Fan, H. Ma, Z. Xu, Performance Analysis of Ship Target Recognition in Multi-Band Infrared Images Based on Deep Learning, *Proceedings of International Conference on Transportation Information and Safety (ICTIS)*, Xian, China, 2023, pp. 432-436. https://doi.org/10.1109/ICTIS60134.2023.10243868

[12] J. Wu, Y. He, J. Zhao, An infrared target images recognition and processing method based on the fuzzy comprehensive evaluation, *IEEE Access*, Vol. 12, pp. 12126-12137, January, 2024. https://doi.org/10.1109/ACCESS.2024.3355157

[13] A. Mikolajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, *Proceedings of International Interdisciplinary PhD Workshop (IIPhDW)*, Swinoujscie, Poland, 2018, pp. 117-122. https://doi.org/10.1109/IIPHDW.2018.8388338

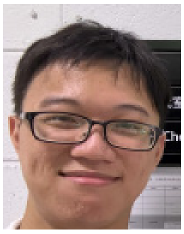
[14] B. M. E. Mahdi, N. Abdelkrim, A. Abdenour, I. Zohir, B. Wassim, D. Fethi, A novel multispectral maritime target classification based on ThermalGAN (RGB-to-thermal image translation), *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 36, No. 8, pp. 1757-1777, November, 2024. https://doi.org/10.1080/0952813X.2023.2165723

[15] Y. M. Zhang, R. Li, Z. Du, Q. Ye, A ship detection method in infrared remote sensing images based on image generation and causal inference, *Electronics*, Vol. 13, No. 7, Article No. 1293, April, 2024. https://doi.org/10.3390/electronics13071293

[16] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, Germany, 2015, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

Biographies

Yan-Ann Chen is currently an assistant professor in the Department of Electronic Engineering, National Taipei University of Technology, Taiwan. He received his Ph.D. in Computer Science from National Chiao Tung University, Taiwan, in 2016. His research interests include Pervasive Intelligence, Edge AI, Multimodal Intelligent Sensing and AIoT.



Hung-Wen Lin received his M.S. degree from Yuan Ze University in 2024. His research and development interests include the Internet of Things and object recognition.